Differential Gene Expression and Mortality Analysis of TCGA-GBM Patients

Brian Zhang, Tina Ryu, Yujung Lin, Johnson Huang

Introduction

In this project, we aimed to discover how certain gene activity, specifically involving the interleukin-13 receptor, is related to the survival of patients with glioblastoma multiforme (GBM). GBM is a very aggressive brain cancer that is quite common and hard to treat effectively because it can appear in different forms, each with a unique genetic blueprint.

Previous research by Han (2018) of The Cancer Genome Atlas (TCGA) data concluded that the gene expressions of interleukin-13 receptors Ra1 and Ra2 are strongly correlated with GBM patient mortality. Our initial analysis results, however, contradicted their findings, showing no significance between IL13Ra1/IL13Ra2 expression and mortality. Given this result, we aimed to find other potential indicators of mortality by expanding our analysis to multigenic expression and different grouping methods.

We used bulk RNA-seq data from TCGA to study genetic factors with computational analysis and machine learning tools. Our computational interests lie in using differential gene expression, mortality analysis, and machine learning methods that can help us find patterns that could tell us which patients are at higher risk, how many days the GBM patients have left, and suggest potential new treatments, like CAR T-cell therapy, might work best. Our goal is to make cancer treatment more personalized, moving away from one-size-fits-all approaches to ones that consider each tumor's unique genetic characteristics.

Methods

The goal of our project was to identify mortality associated genes by developing a differential gene analysis pipeline that models the relationship between patient mortality and expression of immune biomarkers related to glioblastoma multiforme; namely cytokine receptors such as IL4R, IL13Rα1, and IL13Rα2. From our pipeline, bulk RNA-sequencing gene count data from 135 GBM patients were retrieved from the Cancer Genome Atlas (TCGA).

Independent from mortality analysis, since IL4R and IL13Rα receptors were previously examined to study the complex formation during development of GBM tumors, we performed simple linear regression across the gene expression data to determine whether the receptors themselves or receptors in combination with any other genes demonstrated significant correlation in expression level changes (Kawakami et al., 2011).

To analyze the correlation between patient mortality and gene expression, a Kaplan-Meiers analysis was performed. We implemented a function that reads in patient data and outputs the Kaplan-Meier survival curve of the population. At each time point, the survival probability is calculated using the proportion of patients that are still alive. This function allows us to represent patient fate relative to time elapsed since their last surgery; due to the near 100% mortality rate of GBM, interpreting the Kaplan-Meier curve of the patient population using a log-rank test would allow us to conclude if significant difference existed between the patient groups. The log-rank test hypothesizes that the survival rate of two populations are the same. It calculates the expected and observed survival probabilities at each time point and uses chi-square test to obtain the p value in order to determine the significance.

The data were normalized by taking the log2 ratio of expression levels (GBM) against lower grade gliomas (LGGs). Initially, the patients were classified into three groups (lower, same, higher), based on GBM expression levels of genes relative to the expressions of the LGG group. However, due to the lack of public data available, we had to regroup the data into two groups (lower expression ≤ 0 and higher expression > 0 with respect to LGG expression) in order to perform the Kaplan-Meiers survival analysis.

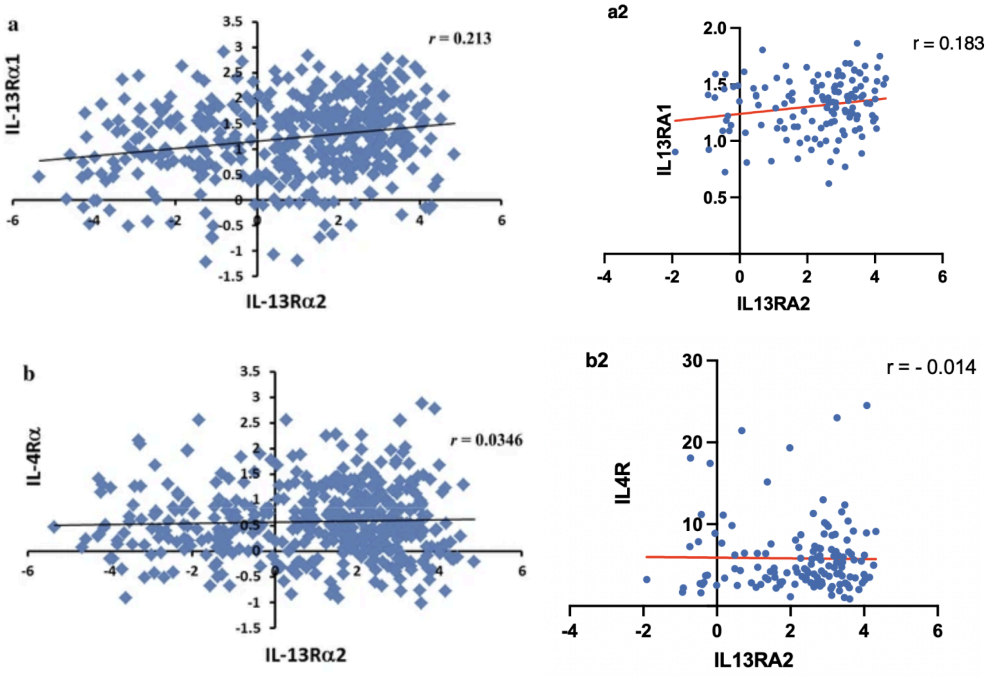
We also performed Pearson and Spearman correlation across the gene expression data and days to death to determine whether IL4R and IL13Rα1 are truly related to mortality rate or if there are other genes that are more closely related to days to death value compared to IL4R and IL13Rα1.

Previous research has illustrated the importance of the relationship between immunomodulatory genes such as IL4R and upstream/downstream signaling molecules such as MYC, SAMSN1, and EGF (Esemen et al., 2022). To further examine this, we implemented a multivariate linear regression model as well as performed principal component analysis on the GBM patient dataset. Multivariate linear regression extends the traditional simple linear regression model by including all the predictor variables (in our case, differentially expressed genes) and controlling for the effects of all predictors on the outcome variable (in our case, the days_to_death variable). The goal of such a model was to estimate the coefficients for each independent variable (individual gene expressions) against the dependent variable (patient mortality) using ordinary least squares, then determine the predictor variables that most greatly affect the response variable. This would allow us to determine a possible set of genes for which overexpression significantly impacts mortality, and perform multigenic survival analysis of these genes against the IL13 receptor complex. We would use cross-validation to examine the accuracy and performance of our model against additional glioma patient data (such as squamous lung carcinoma) and determine the genes for which overexpression significantly impacts mortality rate.

In addition to multivariate linear regression, we implemented a more rigorous recursive feature elimination model; one model based on a Random Forest classifier, and another model based on

XGBoost (Extreme Gradient Boosting). Random Forest is an ensemble learning method that constructs a set of decision trees; each decision tree makes a decision on the outcome variable based on a theoretical set of thresholds, and all the results are aggregated to prevent decorrelation of trees. XGBoost is an extension of this model that iteratively builds on the errors of previous models (in this case previous Random Forest models) until the model has been optimized. Both these methods can be used alongside recursive feature elimination, which simply looks at the variable weightings and coefficients generated by Random Forest and XGBoost and continually removing variables deemed "unimportant" due to a low weighting until a certain set of features has been reached. Using these models, we aimed to identify a set of genes that were implicated in increased GBM patient mortality rate and whose joint overexpression significantly impacted patient outcomes.

Results



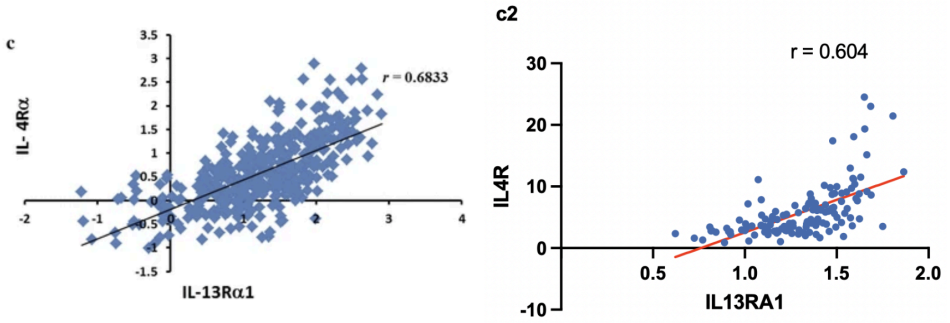


Figure 1. Comparison of correlation between IL-13 receptors expressions in Han et al. and our analysis

correlation between **a, a2** IL-13Ra2 and IL-13Ra1 **b, b2** IL-13Ra2 and IL-4Ra **c, c2** IL-13Ra1 and IL-4Ra, where figures on the left are from Han et al. and on the right are from our analysis. The values are the log of corresponding expression counts.

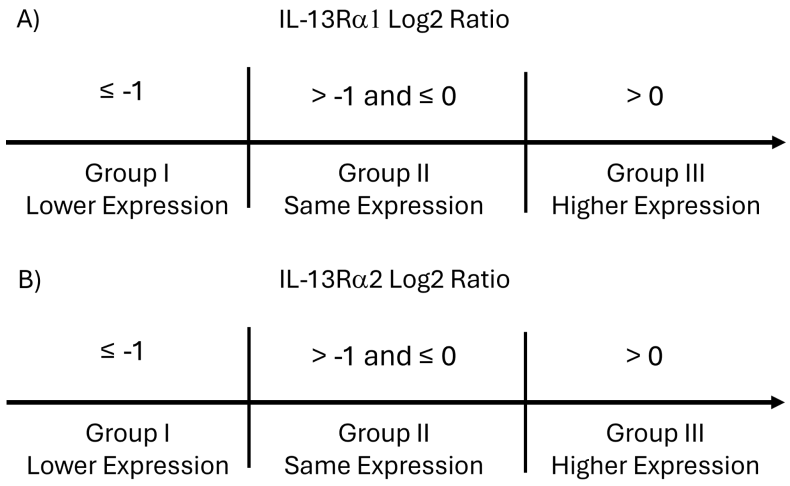


Figure 2. Groups of IL-13Rα1 and IL-13Rα2

The grouping is based on the results of the log2 ratio of each gene expression. Group I is classified as less expression than LGG and contains data with log2 ratio ≤ -1 (n = 21). Group II is classified as the same expression as LGG and contains data with log2 ratio > -1 and ≤ 0 (n = 52). Group III is classified as higher expression than LGG and contains data with log2 ratio > 0. A) the grouping of IL-13Rα1 based on the log2 ratio (n = 63). B) the grouping of IL-13Rα2 based on the log2 ratio. Group I n = 43. Group II n = 16. Group III n =77 .

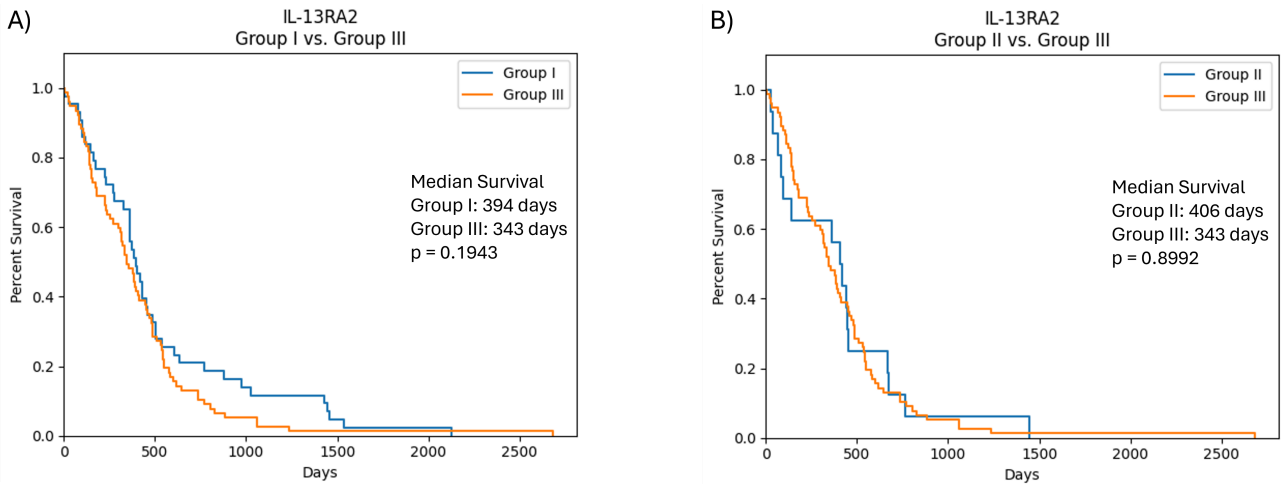


Figure 3. Survival Curve Comparison between groups of IL-13Rα2

The survival curves comparison between 3 groups of IL-13Rα2. The survival curve is generated based on Kaplan Meier Survival Estimate. A) Comparison between Group I and Group III. Median survival day of Group I is 394 days and that of Group III is 343 days. p value of log-rank test between Group I and Group III is 0.1943 (> 0.05) B) Comparison between Group II and Group III. Median survival day of Group II is 406 days and that of Group III is 343 days. p value of log-rank test between Group II and Group III is 0.8992 (> 0.05)

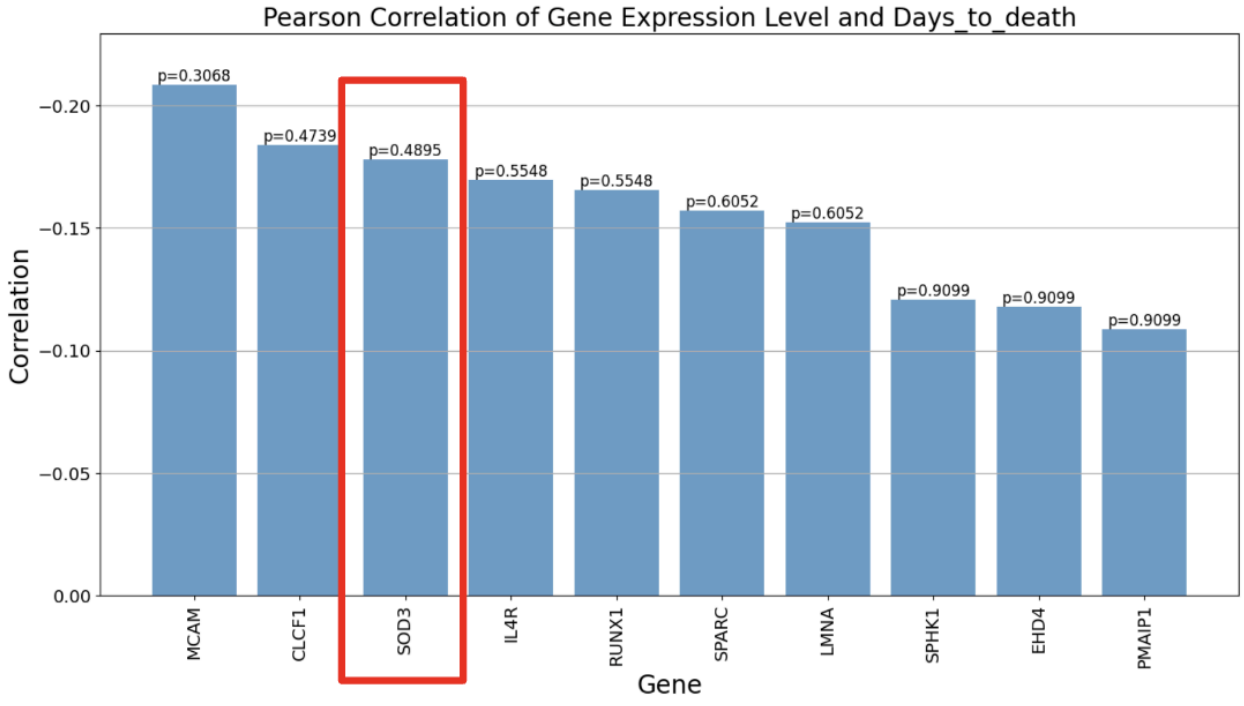


Figure 4. Top 10 related genes to mortality

The top 10 genes negatively correlated with 'days_to_death,' where a more negative correlation coefficient indicates as gene expression increases, 'days_to_death' tends to decrease. All listed p-values are processed under FDR correction, none of the genes are significantly correlated with the mortality that we define. Aside from that, the rank of SOD3 shows that this gene is comparably important to other genes.

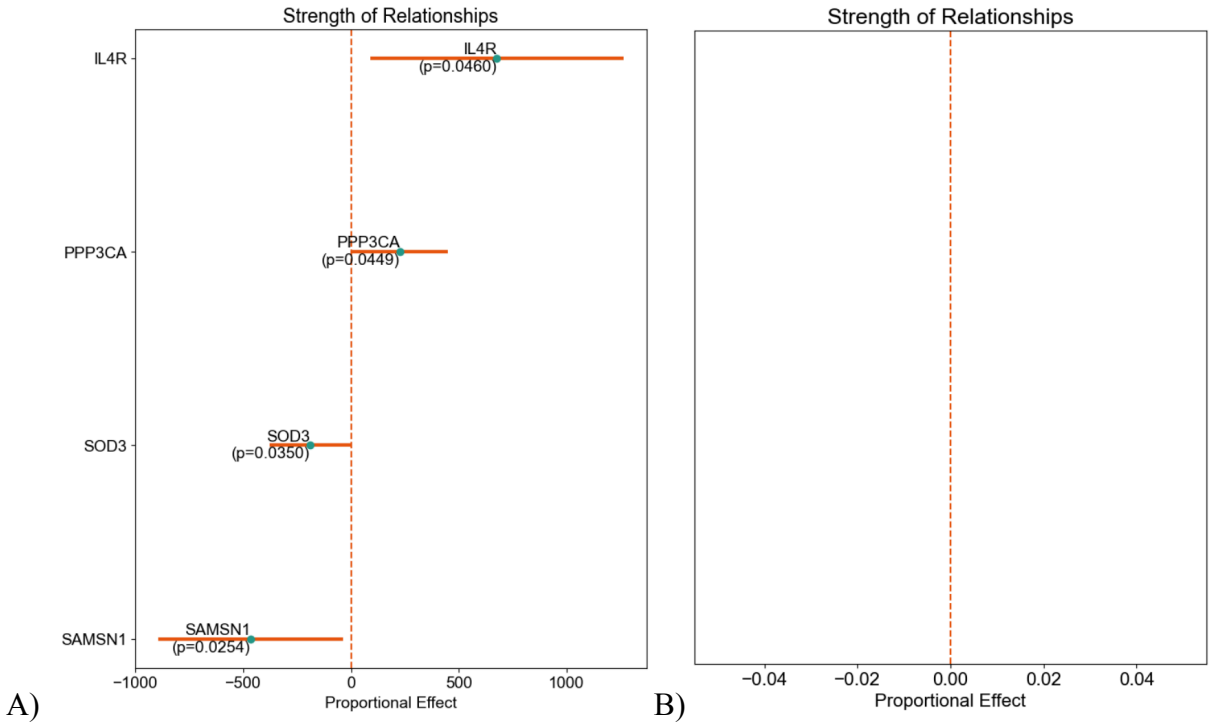
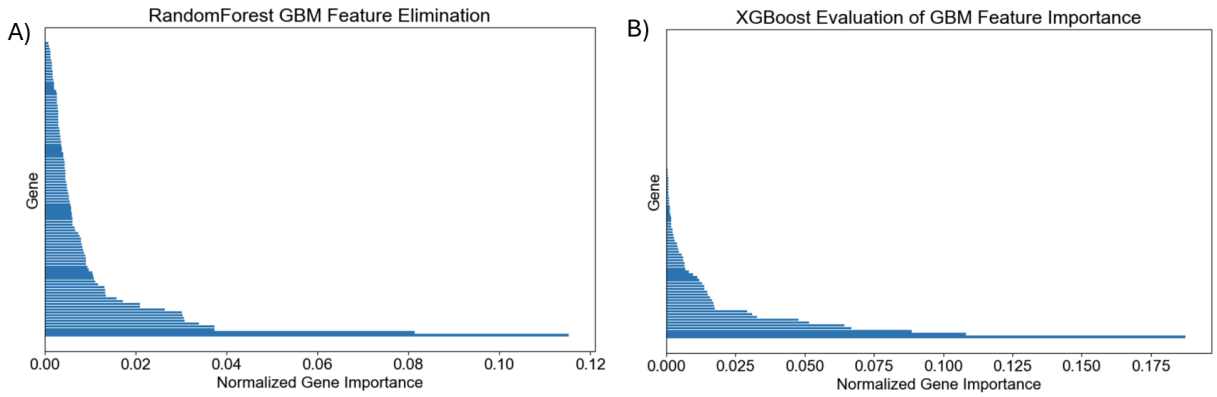


Figure 5. Multivariate Linear Regression Results

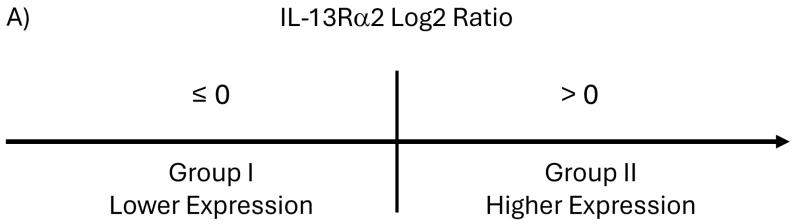
The genes that have a significant proportional effect on the "days_to_death" metric (evaluating mortality rate) are visualized above. A) Four strong predictor genes are identified prior to applying Bonferroni Correction. B) No genes significantly impact mortality following Bonferroni Correction of multivariate linear regression results.



C)

Feature	Importance
SOD3	0.18747
TCF3	0.10830
TDG	0.05392
TLN1	0.051615
RBBP4	0.034522
SNRNP40	0.034219
TTC25	0.031408
NES	0.02445
CDC6	0.02281
CLN3	0.01613
CHD1	0.01206
GART	0.009925
SYNCRIP	0.008058
YBX1	0.00655

Figure 6. Feature Elimination Identifies a Set of Important Genes for Mortality Analysis. A) Identification of feature importance coefficients using a Random Forest classification model. B) Identification of feature importance coefficients using an XGBoost classification model. C) Merged feature importance analysis from Random Forest and XGBoost; the top 15 features identified from recursive feature elimination are listed. Both Random Forest and XGBoost identify SOD3 as the most important gene for mortality analysis.



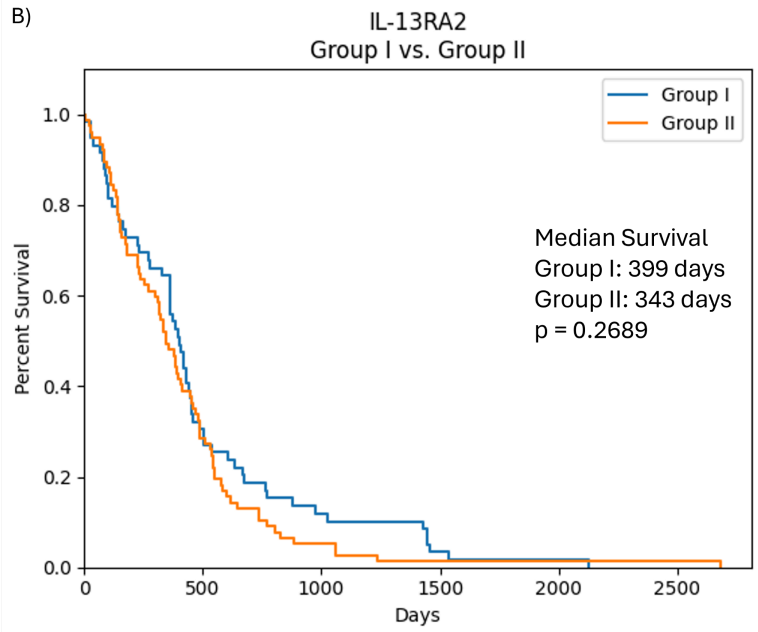
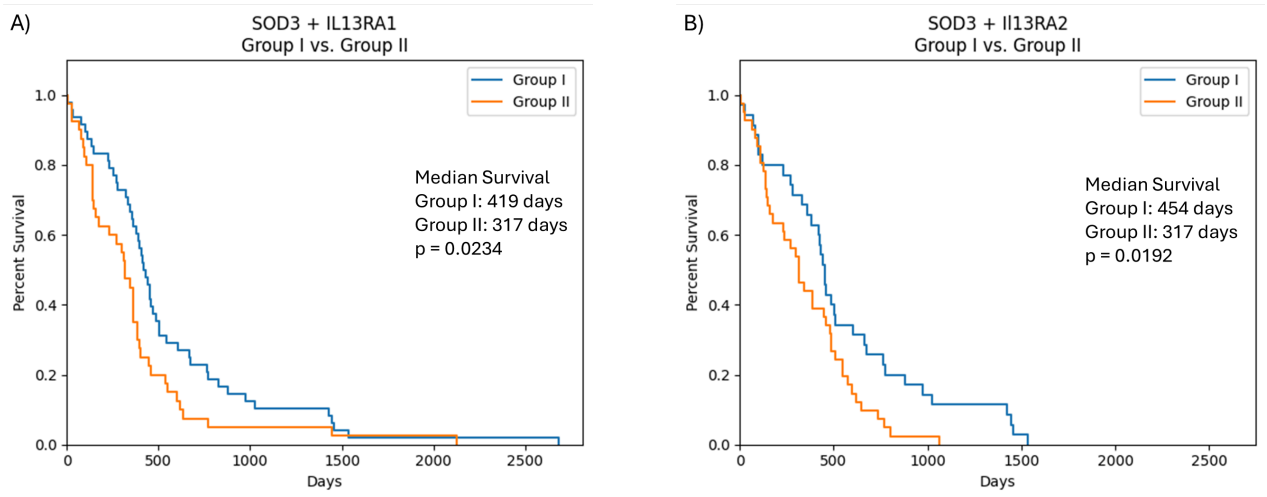


Figure 7. Grouping of IL-13Rα2 and survival curve comparison between Group I and Group II

A) IL-13Rα2 is grouped into two groups based on the log2 ratio. Group I is a lower expression than LGG with log2 ratio less than or equal to 0 (n = 59). Group II is a higher expression than LGG. (n = 77). B) Comparison between Group I and Group II. Median survival day of Group I is 399 days and that of Group II is 343 days. p value of log-rank test between Group I and Group II is 0.2689 (> 0.05).



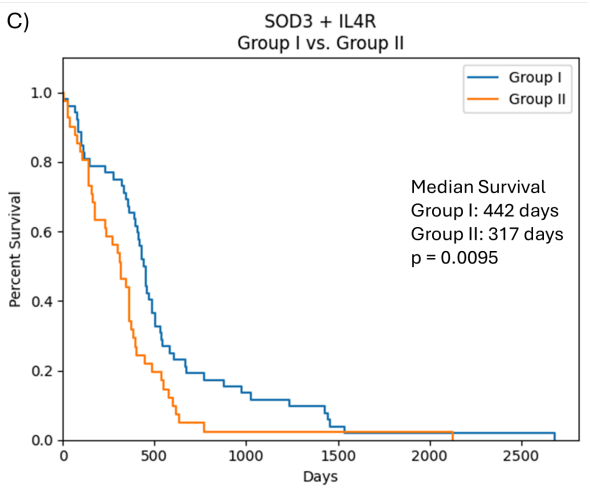


Figure 8. Multigenic Kaplan-Meier survival analysis of Group I and Group II of SOD3 in combination with IL13 complex

Multigenic survival curve comparison of Group I and Group II of SOD3 in combination with IL-13Rα1, IL-13Rα2, and IL-4R. The survival curve is generated based on the multigenic Kaplan-Meier survival estimate and log-rank test. A) Group I and Group II survival curve comparison of SOD3 in combination with IL-13Rα1. Median survival day of Group I is 419 days and 317 days for Group II. p value is 0.0234 (< 0.05). B) Group I and Group II survival curve comparison of SOD3 in combination with IL-13Rα2. Median survival day of Group I is 454 days and 317 days for Group II. p value is 0.0192 (< 0.05). C) Group I and Group II survival curve comparison of SOD3 in combination with IL-4R. Median survival day of Group I is 442 days and 317 days for Group II. p value is 0.0095 (< 0.05).

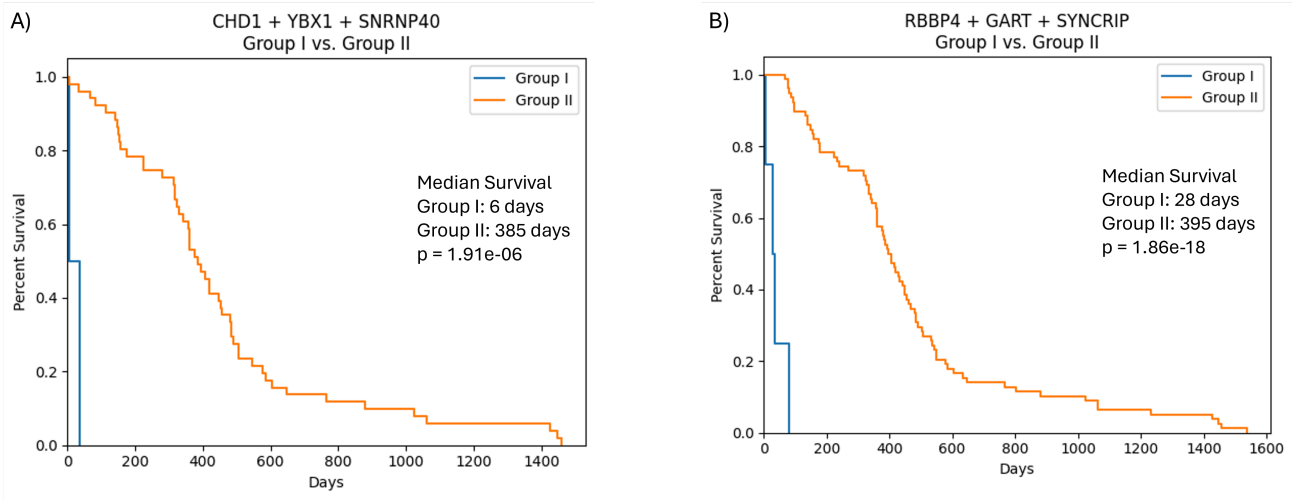


Figure 9. Multigenic Kaplan-Meier survival analysis of two combinations of genes (A) CHD1, YBX1, and SNRNP40, (B) RBBP4, GART, SYNCRIP

Multigenic survival curve comparison of Group I and Group II of combination of three genes. The comparisons are generated based on the multigenic Kaplan-Meier survival estimate and log-rank test. A) Group I and Group II survival curve comparison of CHD1, YBX1, and SNRNP40. Median survival day of Group I is 6 days and 385 days for Group II. p value is 1.91e-06 (< 0.05) following Bonferroni Correction. B) Group I and Group II survival curve comparison of RBBP4, GART, SYNCRIP. Median survival day of Group I is 28 days and 395 days for Group II. p value is 1.86e-18 (< 0.05) following Bonferroni Correction.

Discussion

Our bioinformatics project represents a simple differential gene expression and mortality analysis pipeline to assess relationships between gene expression and surgical outcomes. Using Kaplan-Meier estimators and multivariate linear regression, we are able to model the correlation between different expression groups of IL-13R complex genes and patient survival.

Prior to the mortality analysis, the correlation between expression of different receptor subunits of interleukin 13 were studied, shown in Figure 1. Results show no correlation between IL-13Ra2 and IL-13Ra1 or IL-13Ra2 and IL-4Ra and moderate correlation between IL-13Ra1 and IL-4Ra. Previous studies indicate that there are two types of IL-13 receptors: type I consisting of IL13-Ra1, IL13-Ra2, and IL-4R and type II consisting of Il-4R and IL13-Ra1. The correlation between IL-13Ra1 and IL-4Ra suggests that the receptor complex driving IL-13 signaling in GBM subjects is the type II complex, which is consistent with the results of Han et al's study. Irrespective of the significance of the receptor expression to mortality, this conclusion provides insight into the biological framework of the interleukin-13 signaling pathway in GBM and how it plays a role in the mediation of immune response.

From Figure 3, we can see that 57% of the patients who belonging to group III (higher expression) for IL13Rα2 expression demonstrate a lower expected survival time compared to patients belonging to group I (lower expression) and group II (same expression as LGG)(Figure 2). The plots are similar to the ones from Han's paper. However, the log-rank test results of Group I vs Group III and Group II vs Group III both suggest that there is no significant difference between the survival rate of groups. The results contradict with the conclusions Han made, which suggested that there is an inverse relationship between the expression level of IL-13Rα2 and patient survival.

The correlation analysis from Figure 4 indicates none of the genes in our genes expression data is significantly and negatively correlated to days to death after FDR correction. Similarly in Figure 5, our multivariate linear regression analysis revealed no genes significantly impacting patient mortality following Bonferroni Correction. Due to the fact that multivariate linear regression relies on the assumption of multicollinearity (conditional independence between predictor variables) which may not be the case for our genes selected, we furthered this analysis by performing recursive feature elimination using both a Random Forest and XGBoost classification model. The results from our model, as shown in Figure 6, yielded a set of fifteen genes that were deemed important for predicting mortality.

We suspected the contradicting results were because of the lower number of data in each group due to the lack of publicly available data. Therefore, we regrouped the data into two groups and performed the Kaplan-Meier survival analysis and log-rank test again (Figure 7). Group I consisted of data with log2 ratio less than or equal to zero and was classified as lower expression than LGG. Group II consisted of data with log2 ratio larger than zero and was classified as higher expression than LGG. However, the results following regrouping still suggested there is no significant difference between the mortality rates of patients with lower expression and higher expression of IL-13Rα2 (Figure 7). We then decided to use the results of our feature elimination model to perform multigenic survival analysis. By looking at the combined expression pattern of SOD3 and the IL13 complex as identified by RFE, we demonstrated that the joint overexpression of SOD3 and IL13 receptor complex genes is significantly implicated in increased patient mortality (Figure 8). SOD3 itself is a protein-encoding gene that encodes for a superoxide dismutase, which traditionally protects the brain from oxidative stress. However, recent studies have identified variants in SOD3 that are overexpressed in brain tumors and may play roles in tumor progression (Wert et al., 2018). Looking at additional combinations of genes identified by RFE, we identified two additional multigene groups: a group consisting of CHD1, YBX1, SNRNP40, and a group consisting of RBBP4, GART, and SYNCRIP. Performing multigenic survival analysis on these two groups showed a highly significant correlation between group joint overexpression and reduced patient mortality (Figure 9). Previous studies examining these genes separately have identified these genes within other human cancers such as prostate cancer or liver cancer, where they have been shown to express tumor-suppressive capabilities (Li et al., 2023). Our results indicate a possible extension of these functions within glioblastoma

multiforme tumors, though additional experimentation is necessary to robustly understand this relationship.

Given the results of our pipeline, it would be interesting to further our analysis through gene expression - chemotherapy resistance association studies; for instance, attempting to establish a link between overexpression of IL-13Rα2 and resistance to methotrexate immunosuppression during chemotherapy. Due to the lack of publicly available chemotherapy treatment data online, we were unable to replicate the results of the original paper in this regard, which found a significant link between overexpression of IL-13Rα2 and resistance to temozolomide chemotherapy. In addition to this, due to the reliance of mortality analysis on data quality and reliability, it may be necessary to repeat our analyses with more rigorous experimental approaches. For instance, it would be interesting to see if our results hold true for data collected from a longitudinal analysis of GBM patients, or following a functional enrichment analysis (eg: gene ontology) to determine biological mechanisms of multigenic IL-13 receptor function.

Conclusions

In conclusion, our project aims to evaluate the relationship between different expression patterns of groups of related genes to the IL-13 receptor complex and patients with glioblastoma multiforme. By implementing a differential gene analysis pipeline, we determined that, in contrast to the original paper findings by Han et al., there does not exist a significant relationship between IL13Rα1 and IL13Rα2 expression levels and poor GBM patient prognosis. However, multigenic survival curve expression analysis with additional genes of interest reveals a significant relationship between the IL-13 receptor complex and SOD3 expression levels and reduced patient mortality, suggesting the importance of SOD3 as a potential therapeutic target for IL-13 complex regulation or as a biomarker for GBM disease outcome. In addition, multigenic Kaplan-Meier survival analysis identified two clusters of genes CHD1+YBX1+SNRNP40 and RBBP4+GART+SYNCRIP, where its combined expression positively correlated with reduced patient mortality. These identified clusters can be of interest for further analysis for use in patient outcome biomarkers..

References

1. Esemen, Y., Awan, M., Parwez, R., Baig, A., Rahman, S., Masala, I., Franchini, S., & Giakoumettis, D. (2022). Molecular Pathogenesis of Glioblastoma in Adults and Future Perspectives: A Systematic Review. International journal of molecular sciences, 23(5), 2607. https://doi.org/10.3390/ijms23052607
2. Han, J., Puri, R.K. Analysis of the cancer genome atlas (TCGA) database identifies an inverse relationship between interleukin-13 receptor α1 and α2 gene expression and poor prognosis and drug resistance in subjects with glioblastoma multiforme. *J Neurooncol* 136, 463–474 (2018). https://doi.org/10.1007/s11060-017-2680-9
3. Li, H., Gigi, L., & Zhao, D. (2023). CHD1, a multifaceted epigenetic remodeler in prostate cancer. Frontiers in oncology, 13, 1123362. https://doi.org/10.3389/fonc.2023.1123362

4. Kawakami K, Taguchi J, Murata T, Puri RK. The interlukin-13 receptor α2 chain: an essential component for binding and internalization but not for interlukin-13-induced signal transduction through the STAT6 pathway. Blood. 2011;97(9):2673–2679. doi: 10.1182/blood.V97.9.2673.
5. Thaci B, Brown CE, Binello E, Werbaneth K, Sampath P, Sengupta S. Significance of interleukin-13 receptor alpha 2-targeted glioblastoma therapy. Neuro Oncol. 2014;16(10):1304–1312. doi: 10.1093/neuonc/nou045.
6. Wert, K. J., Velez, G., Cross, M. R., Wagner, B. A., Teoh-Fitzgerald, M. L., Buettner, G. R., McAnany, J. J., Olivier, A., Tsang, S. H., Harper, M. M., Domann, F. E., Bassuk, A. G., & Mahajan, V. B. (2018). Extracellular superoxide dismutase (SOD3) regulates oxidative stress at the vitreoretinal interface. Free radical biology & medicine, 124, 408–419. https://doi.org/10.1016/j.freeradbiomed.2018.06.024